

Mixture Model Averaging for Clustering and Classification

Yuhong Wei and Paul D. McNicholas*

Department of Mathematics & Statistics, University of Guelph.

Abstract

In mixture model-based clustering applications, it is common to fit several models from a family and report clustering results from the ‘best’ one. Selection of this best model is a difficult and consequential problem and criteria commonly used include the Bayesian information criterion, the Akaike information criterion, and the integrated completed likelihood. We propose an alternative to the selection of a best model, instead averaging the clustering results of several models. In the course of model averaging, the top few models often have different numbers of mixture components and so merging components is necessary. The effectiveness of our model-based clustering averaging approach is illustrated using a family of Gaussian mixture models on simulated and real data. This paper is perhaps the first step in a departure from the ‘single best model’ paradigm that currently dominates the model-based clustering literature.

1 Introduction

Model-based clustering is an idiom that describes the application of a mixture model, or any model, for clustering. Dating at least as far back as Wolfe (1963), interest in model-based clustering is increasing steadily in application areas like food authenticity, social networks, and microarray gene expression analysis (see Yeung et al., 2001; Wehrens et al., 2004; Krivitsky et al., 2009; McNicholas and Murphy, 2010; Murphy et al., 2010, for examples). A random vector \mathbf{X} arises from a parametric finite mixture distribution if its density can be written $f(\mathbf{x}|\boldsymbol{\vartheta}) = \sum_{g=1}^G \pi_g f_g(\mathbf{x}|\boldsymbol{\theta}_g)$, where G is the number of components, π_g are mixing proportions ($\pi_g > 0$, $\sum_{g=1}^G \pi_g = 1$), and $\boldsymbol{\vartheta} = (\pi_1, \dots, \pi_G, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_G)$ denotes the parameters. Gaussian mixtures have dominated the model-based clustering literature until very recently. The likelihood for p -dimensional $\mathbf{x}_1, \dots, \mathbf{x}_n$ from a Gaussian mixture model is

$$\mathcal{L}(\boldsymbol{\vartheta}) = \prod_{i=1}^n \sum_{g=1}^G \pi_g \phi(\mathbf{x}_i | \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \quad (1)$$

where $\phi(\mathbf{x}_i | \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ is the density of a multivariate Gaussian distribution with mean $\boldsymbol{\mu}_g$ and covariance $\boldsymbol{\Sigma}_g$, and $\boldsymbol{\vartheta}$ once again denotes the model parameters.

The density in Equation 1 has a total of $(G-1) + Gp + Gp(p+1)/2$ free parameters and so it is practically necessary to introduce parsimony. A total of $Gp(p+1)/2$ of these parameters are in the component covariances and so imposing the isotropic constraint on the component covariances, i.e., $\boldsymbol{\Sigma}_g = \sigma_g \mathbf{I}_p$, is a very simple way to reduce the number of parameters from quadratic to linear in p . Of course, this constraint will not be practical for most applications and so less restrictive constraints need to be considered. Several such approaches have been tried, usually based on imposing constraints upon a decomposed covariance structure.

*Department of Mathematics & Statistics, University of Guelph, Guelph, Ontario, N1G 2W1, Canada. E-mail: paul.mcnicholas@uoguelph.ca.

The most famous such approach is based on an eigen-decomposition (Banfield and Raftery, 1993), so that $\Sigma_g = \lambda_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}_g'$, where \mathbf{D}_g is the orthogonal matrix of eigenvectors of Σ_g , \mathbf{A}_g is a diagonal matrix with elements proportional to its eigenvalues, and λ_g is a scalar. Celeux and Govaert (1995) impose constraints on λ_g , \mathbf{D}_g , and \mathbf{A}_g to obtain a family of 14 mixture models. The famous MCLUST family of models (Table 1) is the subset of these 14 that is supported by the `mclust` package (Fraley and Raftery, 2006) for R (R Development Core Team, 2012). Parameter estimation for each member of the family is carried out using the expectation-maximization (EM) algorithm (Dempster et al., 1977); extensive details on the EM algorithm and applications to mixture models are given by McLachlan and Krishnan (2008).

Table 1: Models available in the `mclust` package, along with number of covariance parameters in each case.

	Dist.	Volume	Shape	Orient.	Σ_g	Free cov. paras.
EII	Spherical	Equal	Equal	NA	$\lambda \mathbf{I}$	1
VII	Spherical	Variable	Equal	NA	$\lambda_g \mathbf{I}$	G
EEI	Diagonal	Equal	Equal	Cord. Axes	$\lambda \mathbf{A}$	p
VEI	Diagonal	Variable	Equal	Cord. Axes	$\lambda_g \mathbf{A}$	$p + G - 1$
EVI	Diagonal	Equal	Variable	Cord. Axes	$\lambda \mathbf{A}_g$	$Gp - G + 1$
VVI	Diagonal	Variable	Variable	Cord. Axes	$\lambda_g \mathbf{A}_g$	Gp
EEE	Ellipsoidal	Equal	Equal	Equal	$\lambda \mathbf{D} \mathbf{A} \mathbf{A}'$	$p(p+1)/2$
EEV	Ellipsoidal	Equal	Equal	Variable	$\lambda \mathbf{D}_g \mathbf{A} \mathbf{A}_g'$	$Gp(p+1)/2 - (G-1)p$
VEV	Ellipsoidal	Variable	Equal	Variable	$\lambda_g \mathbf{D}_g \mathbf{A} \mathbf{A}_g'$	$Gp(p+1)/2 - (G-1)(p-1)$
VVV	Ellipsoidal	Variable	Variable	Variable	$\lambda_g \mathbf{D}_g \mathbf{A}_g \mathbf{A}_g'$	$Gp(p+1)/2$

A typical application of the MCLUST family of models consists of running each of the ten models (Table 1) for a range of values of G ; the default in `mclust` is $G = 1, \dots, 9$, which results in 90 models being fitted. As is typical with families of mixture models, the best of these 90 models will be selected using some criterion and the associated classifications reported. The most popular criterion for this purpose is the Bayesian information criterion (BIC; Schwarz, 1978), $\text{BIC} = 2l(\mathbf{x}, \hat{\boldsymbol{\vartheta}}) - \rho \log n$, where $\hat{\boldsymbol{\vartheta}}$ is the maximum likelihood estimate of $\boldsymbol{\vartheta}$, $l(\mathbf{x}, \hat{\boldsymbol{\vartheta}})$ is the maximized log-likelihood, and ρ is the number of free parameters. Leroux (1992) and Keribin (2000) give theoretical results that, under certain regularity conditions, support the use of BIC for choosing the number of components in a mixture model. In addition, in a range of applications of model-based clustering, model selection based on the BIC has given good results (cf. Kass and Wasserman, 1995; Dasgupta and Raftery, 1998; Fraley and Raftery, 1998; Stanford and Raftery, 2000). None the less, it is well known that the model selected based on BIC does not necessarily give the best predicted classifications or the most accurate number of components.

Several alternatives to the BIC have been proposed. The Akaike information criterion (AIC), which was considered in the mixture modelling context by Bozdogan and Sclove (1984), takes the form $\text{AIC} = 2l(\mathbf{x}, \hat{\boldsymbol{\vartheta}}) - 2\rho$. However, it has been observed that the AIC is order inconsistent and tends to over-fit models (see Koehler and Murphree, 1988, for example). Biernacki et al. (2000) proposed the integrated completed likelihood (ICL), which is approximated by penalizing the BIC using the estimated mean entropy. We have $\text{ICL} \approx \text{BIC} + \sum_{i=1}^N \sum_{g=1}^G \text{MAP} \{\hat{z}_{ig}\} \log \hat{z}_{ig}$, where $z_{ig} = 1$ if \mathbf{x}_i belongs to the g th component and $z_{ig} = 0$ otherwise, and MAP denotes the maximum *a posteriori* probability such that $\text{MAP} \{\hat{z}_{ig}\} = 1$ if $\max_g \{\hat{z}_{ig}\}$ occurs in the g th component and $\text{MAP} \{\hat{z}_{ig}\} = 0$ otherwise. The ICL is an attempt to shift the focus from selection of the correct number of components G to selection of the right number of clusters. Another way to do this is to use the BIC, or another criterion, for model selection and then to merge the resulting Gaussian components to give clusters (cf. Hennig, 2010). Li (2005) proposed fitting a multilayer mixture (i.e., a mixture of Gaussian mixtures), assuming that the number of clusters k is known in advance, and then applies k -means clustering to the G component means, where G is the estimated number of Gaussian mixture components based on the BIC. However, the method suggested by Li (2005) has the drawback that it assumes *a priori* knowledge of the number of clusters and it does not meaningfully take account of

cluster shapes. Baudry et al. (2008) proposed fitting a Gaussian mixture model to data, then selecting the total number of Gaussian mixture components using BIC, and combining them hierarchically according to an entropy criterion.

The goal of this paper is to depart from the paradigm wherein a single best model is selected. Instead, we report clustering results based on an averaging of the top few models. The remainder of this paper is outlined as follows. In Section 2, we describe our model averaging schemata, before illustrating our approach on real data (Section 3). The paper concludes with discussion and suggestions for future work (Section 4).

2 Methodology

2.1 Overview

As mentioned earlier, it is common to fit many mixture models within a family and then report clustering results based only on the best one. Such criterion-based methods of model selection have the general feature that the larger the value of the criterion, the stronger the ‘evidence’ for the model (i.e., covariance structure and number of components, as well as number of latent variables where relevant). We argue that criterion-based approaches are not reasonable because they are effectively throwing away all but the best model. One may argue, inferentially, that it has to matter that 90 (say) models were fitted when reporting results based on just one. Practically, criterion-based approaches can be considered unreliable when the difference between the largest value of a criterion is close to other values. To overcome this problem, we consider a model averaging approach. More specifically, we consider an averaging of the top few mixture models with applications in clustering and classification. Herein, we consider cases where the number of models to be averaged is allowed to vary and cases where it is fixed *a priori*.

2.2 Top Model Selection

Herein, a top models set S is defined and will contain the ‘top models’ to be averaged. The selection procedure for the top models is the same for each of the three criteria, and in what follows here the BIC is utilized for illustration. Prior to the selection of the elements of S , the top models are ordered according to their BIC values, e.g., denote the largest BIC value by $\text{BIC}[1]$, and the corresponding model is the 1st model; the second largest BIC value is $\text{BIC}[2]$, and the model is the 2nd model; likewise for the other models. Two scenarios are now considered: selection of the ‘top M ’ and ‘top 5’ models, respectively.

Top M Model Selection

1. Step 1: Select the 1st model, and put it into S . Set $k = 2$,
2. Step 2: Select the k th model, and calculate $\text{DIFF} = \text{BIC}[1] - \text{BIC}[k]$.
3. Step 3: If $\text{DIFF} < 10$, put the k th model into S , set $k \leftarrow k + 1$, and return to Step 2. Else terminate.

The resulting set S contains M models, i.e., the top M models.

Top 5 Model Selection

1. Select the top 5 models (corresponding to the first five highest value of the BIC), and put them into set S .

The resulting set S contains 5 models, i.e., the top 5 models.

2.3 Merging Mixture Components

If models within the set S have different numbers of components, we will need to merge the Gaussian mixture components. In this section, we introduce a procedure for merging a larger number of mixture components into a smaller number of clusters. For the model with larger number of components, e.g. G , and the new model after the merging process has smaller number of clusters, e.g. H , where $H < G$, the density takes the form $f(\mathbf{x}) = \sum_{j=1}^H \pi_j^* f_j^*(\mathbf{x}) = \sum_{g=1}^G \pi_g \phi(\mathbf{x} | \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$, where π_j^* is the sum of π_g of the Gaussian mixture components assigned to f_j^* , and $f_j^*(\mathbf{x})$ is from a single or a mixture of component densities from the original model, such that each original Gaussian component appears in exactly one of $f_j^*(\mathbf{x}), \dots, f_k^*(\mathbf{x})$. Suppose, for example, that $G = 3$ and $H = 2$. Then $f_1^*(\mathbf{x})$ could be a mixture of the first two Gaussian components and $f_2^*(\mathbf{x})$ would equal the third one.

We use the adjusted Rand index (ARI; Rand, 1971; Hubert and Arabie, 1985) in our merging criteria. The ARI is the Rand index corrected for chance. The Rand index compares two data partitions and is given by the number of pair agreements divided by the total number of pairs. The correction leading to the ARI is performed to account for the fact that if classification is performed randomly then some cases will be correctly classified by chance.

The same notation from Section 2.2 is used here, where BIC[1] refers to the largest BIC value, and the corresponding model is referred as the 1st model, etc. We merge components to make the other models in S comparable to a ‘reference model’. Two cases are considered.

- In Case I, the 1st model has G components and is the reference model. All remaining models in set S are to be merged to G clusters.
- In Case II, the i th ($i \neq 1$) model has the least number of components G and is the reference model. The remaining models (including the 1st model) are merged to have G components.

Note that we are assuming that the BIC will only overestimate the number of components; therefore, we ignore models with less components than the reference model.

The merging process is best illustrated with an example. Consider merging 7 components into 4 components, where 7 is the number of components from a model in S and 4 is the least number of components in S . Suppose we work under Case II, so the reference model has $G = 4$ components. The partition corresponding to the reference model is used as the underlying true classification in the merging process and is called the ‘reference partition’. In what follows here, to avoid confusion, we use $\{1, 2, \dots, 7\}$ to denote components of the seven-component model and $\{a, b, c, d\}$ to denote components in the $G = 4$ component reference model. The following steps illustrate our merging strategy.

1. A combination matrix \mathbf{A} of size 35×4 is generated, note that $\binom{7}{4} = 35$, where each row represents a partial clustering, e.g., a sample row $\mathbf{a}_{22} = (2, 3, 4, 6)$ means that group 2 goes into new group a , group 3 goes into new group b , group 4 goes into new group c , and group 6 goes into new group d .
2. For each row in \mathbf{A} , we must deal with the remaining components; e.g., $\{1, 5, 7\}$ in the case of row 22. A permutation matrix \mathbf{B} (64×3) is created to include all possibilities. One sample row is $\mathbf{b}_j = (a, a, d)$, which puts component 1 into new component a , component 5 into new component a , and component 7 into new component d , so the model after merging has the new groups $\{a, b, c, d\} = \{1 \cup 2 \cup 5, 3, 4, 6 \cup 7\}$.
3. The ARIs between the reference partition and all partitions arising from the model after merging are calculated and recorded in the 35×64 matrix \mathbf{C} . For example, in step 1, $\mathbf{a}_{22} = (2, 3, 4, 6)$ is the 22nd row in matrix \mathbf{A} , in step 2, $\mathbf{b}_j = (a, a, d)$ is the 4th row in matrix \mathbf{B} , and the ARI value between reference partition and this model after merging is stored at the 22th row and the 4th column of the matrix \mathbf{C} .

Once we have every element of the matrix \mathbf{C} , the best merging combination is chosen to correspond to the largest ARI value.

2.4 Top Model Averaging

After merging components where needed so that each model within S has the same number, and discarding those with less components than the reference model (if applicable), it remains to find the estimated conditional probability for the model after merging and then to average these conditional probabilities across the models in S . Denote by Z_{ig}^G the conditional posterior probability that \mathbf{x}_i arises from component g with respect to the G -component solution. If (only) components g' and g'' from the G -component solution are combined or merged into one cluster, then the Z_{ig} remain the same except for g' and g'' , and the new component $g' \cup g''$ has the conditional probability $Z_{ig' \cup g''} = Z_{ig'}^G + Z_{ig''}^G$.

Sticking with the notation from Section 2.2, model averaging is a process to compute the weighted summation of the models in the set S , with $Z_{\text{ave}} = \sum w[i]Z[i]$, where Z_{ave} is the averaged model conditional probability matrix, $w[i]$ is the weight of the i th model defined later, and $Z[i]$ is the conditional probability matrix for the i th model. The purpose of the $w[i]$ is to weight the importance of each model. Essentially, the higher the value of the criterion (BIC, ICL, or AIC), the stronger the evidence for the model and the number of components, and the larger this weight should be. In this regard, $w[i]$ is defined as

$$w[i] = \frac{1/v[i]}{\sum 1/v[j]}, \quad (2)$$

where $v[i]$ is the value of the criterion for the i th model. Note that higher $v[i]$ values produce larger $w[i]$.

3 Data Analyses

3.1 Italian Wine

Forina et al. (1986) present data on chemical and physical properties of 178 samples of three varieties (Barolo, Barbera, and Grignolino) of wine from the Piedmont region of Italy. A data set containing 27 physical and chemical properties is available in the `pgmm` package (McNicholas et al., 2011) for R. We ran `mclust` on these data, under the default settings. The VVI model with $G = 3$ components was the best model in terms of BIC (−12103.76) and the EVI model had almost identical BIC (−12103.82). The ICL chose the same top two models, albeit in the opposite order, and the AIC selected a $G = 8$ component model. We then applied our top 5 and top M averaging procedures (merging components where needed). The results for the top 5 averaging approach (Table 2) show that averaging led to much better clustering performance (taking the wine regions as the correct classes). This is true for the BIC (ARI of 0.914 versus 0.895), the ICL (ARI of 0.914 versus 0.830), and the AIC (0.722 versus 0.386). The results for the top M approach (Table 3) are not as impressive, with averaging performing as well as or worse than selecting the best model. Note that M was small for all three criteria: $M = 2$ for BIC and ICL, and $M = 1$ for AIC.

3.2 Italian Olive Oil

Forina and Tiscornia (1982) present data on the percentage composition of eight fatty acids found by lipid fraction of 572 olive oils from three regions of Italy (Southern Italy, Sardinia, and Northern Italy). These three regions can be further broken down into nine areas. Again, these data are available in the `pgmm` package. As before, we ran `mclust` on these data using the default settings. The BIC and the ICL both selected a VVV model with $G = 5$ components. The AIC selected a VVV model with $G = 9$ components. We then applied our top 5 and top M averaging procedures, merging components where needed. The results for the top 5 averaging approach (Table 4) show that averaging led to superior clustering performance no matter whether the regions or the areas were taken as the true classes. The top M approach could not be applied to these data because the difference between the best and subsequent models was more than 10 under each criterion.

Table 2: Results for top 5 model selection, merging, and averaging for the wine data.

	Model Selection			Merging		Average Weight	ARI	
	Models	Criteria	G	Case I	Case II		Regions	Areas
BIC	1st(VVI)	-12103.76	3	no need		0.2007217	Best:	
	2nd(EVI)	-12103.82	3	no need		0.2007208	0.8950872	
	3rd(VEI)	-12145.07	7	$\{1 \cup 2, 3 \cup 4 \cup 5, 6 \cup 7\}$	no need	0.200039	Averaged:	
	4th(EEI)	-12191.58	5	$\{1, 2 \cup 3, 4 \cup 5\}$		0.199276	0.9135014	
	5th(EEI)	-12193.62	6	$\{1 \cup 2, 3 \cup 4, 5 \cup 6\}$		0.1992425		
ICL	1st(EVI)	-12104.4	3	no need		0.2007521	Best:	
	2nd(VVI)	-12104.45	3	no need		0.2007513	0.8301066	
	3rd(VEI)	-12149.27	7	$\{1 \cup 2, 3 \cup 4 \cup 5, 6 \cup 7\}$	no need	0.2000107	Averaged:	
	4th(EEI)	-12193.59	5	$\{1, 2 \cup 3, 4 \cup 5\}$		0.1992837	0.9135014	
	5th(EEI)	-12198.58	6	$\{1 \cup 2, 3 \cup 4, 5 \cup 6\}$		0.1992022		
AIC	1st(EVI)	-8862.6	8		$\{1 \cup 2 \cup 3, 4, 5 \cup 6, 7, 8\}$	0.2165185	Best:	
	2nd(EVI)	-9152.212	7		$\{1 \cup 2, 3, 4 \cup 5, 6, 7\}$	0.209667	0.3860865	
	3rd(VEI)	-9643.861	7	no need	$\{1 \cup 2, 3, 4 \cup 5, 6, 7\}$	0.198978	Averaged:	
	4th(EEI)	-10052.09	6		$\{1 \cup 2, 3, 4, 5, 6\}$	0.1908972	0.7215702	
	5th(VEV)	-10432.34	5		no need	0.1839393		

Table 3: Results from top M model selection, merging, and averaging for the wine data.

	Model Selection			Merging	Average	ARI
	Models	Criteria	G	Cases I & II	Weight	
BIC	1st(VVI)	-12103.8	3	no need	0.5000012	Best: 0.8950872
	2nd(EVI)	-12103.8	3		0.4999988	Averaged: 0.8461782
ICL	1st(EVI)	-12104.4	3	no need	0.500001	Best: 0.8301066
	2nd(VVI)	-12104.5	3		0.499999	Averaged: 0.8301066
AIC	1st(EEV)	-8862.6	8	no need	no need	Best: 0.3860865

Table 4: Results for top 5 model selection, merging, and averaging for the olive data.

	Model Selection			Merging		Average Weight	ARI	
	Models	Criteria	G	Case I	Case II		Regions	Areas
BIC	1st(VVV)	-5096.72	5	no need		0.2040479	Best:	Best:
	2nd(VVV)	-5166.82	9	$\{1 \cup 2, 3 \cup 4, 5 \cup 6, 7, 8 \cup 9\}$		0.2012789	0.5969466	0.776304
	3rd(VVV)	-5225.58	8	$\{1 \cup 8, 2, 3 \cup 4, 5 \cup 6, 7\}$	no need	0.1990155	Averaged:	Averaged:
	4th(VVV)	-5228.28	6	$\{1, 2 \cup 3, 4, 5, 6\}$		0.1989128	0.6057673	0.8056072
	5th(VVV)	-5285.89	7	$\{1, 2 \cup 3, 4 \cup 5, 6, 7\}$		0.1967449		
ICL	1st(VVV)	-5098.71	5	no need		0.2044837	Best:	Best:
	2nd(VVV)	-5177.5	9	$\{1 \cup 2, 3 \cup 4, 5 \cup 6, 7, 8 \cup 9\}$		0.2013719	0.5969466	0.776304
	3rd(VVV)	-5235.94	8	$\{1 \cup 8, 2, 3 \cup 4, 5 \cup 6, 7\}$	no need	0.1991243	Averaged:	Averaged:
	4th(VVV)	-5250.09	6	$\{1, 2 \cup 3, 4, 5, 6\}$		0.1985875	0.6057673	0.8056072
	5th(VVV)	-5307.68	7	$\{1, 2 \cup 3, 4 \cup 5, 6, 7\}$		0.1964326		
AIC	1st(VVV)	-3409.77	9		$\{1 \cup 2, 3 \cup 4, 5 \cup 6, 7, 8 \cup 9\}$	0.2238324	Best:	Best:
	2nd(VVV)	-3664.24	8		$\{1 \cup 8, 2, 3 \cup 4, 5 \cup 6, 7\}$	0.2082877	0.3437476	0.6600636
	3rd(VVV)	-3920.26	7	no need	$\{1, 2 \cup 3, 4 \cup 5, 6, 7\}$	0.1946852	Averaged:	Averaged:
	4th(VVV)	-4058.36	6		$\{1, 2 \cup 3, 4, 5, 6\}$	0.1880603	0.6057673	0.8056072
	5th(VVV)	-4122.5	5		no need	0.1851345		

3.3 Simulated Data

Data were simulated from a mixture of $G = 4$ multivariate Gaussian distributions using the `genRandomClust()` function from the R package `clusterGeneration` (Qiu and Joe., 2012). The `genRandomClust()` function generates random clusters based on the method proposed by Qiu and Joe (2006). We generate $p = 3$ variables for $n = 562$ observations and we choose to generate substantially overlapping clusters (Figure 1). This is a very difficult clustering problem and perfect classification is not expected. We ran `mclust` on these data, under the default settings. The EII model with $G = 5$ was selected by the BIC, the ICL selected a $G = 4$ component EVI model, and the AIC chose a $G = 9$ component EVI model.

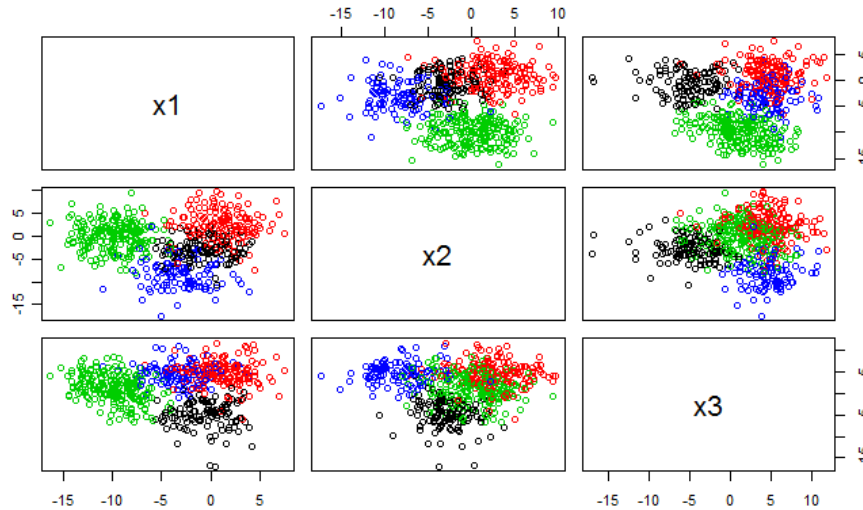


Figure 1: Pairs plot of the $G = 4$ component, $p = 3$ variable data generated by `genRandomClust()`.

We then applied our top 5 and top M averaging procedures, merging components where needed. The results for the top 5 averaging approach (Table 5) show that averaging led to far superior clustering performance when the BIC is used, taking the ARI from 0.714 for the single best model to 0.823 (Case I) or 0.878 (Case II). Top 5 merging had no effect for the ICL but did increase the ARI when the AIC was used (from 0.535 to 0.593). The results for the top M approach (Table 3) are not applicable for the BIC or the ICL but had a dramatic effect when used with the AIC. A total of 17 models had an AIC value within 10 of the best model and averaging these 18 clustering models, with Case II merging, resulted in a remarkably improved ARI (0.881 versus 0.535).

4 Summary and Discussion

This paper is the first step in what will perhaps ultimately be a departure from the ‘single best model’ paradigm that has heretofore dominated the model-based clustering literature. The averaging approaches used are not technically complicated but we show that they can produce extraordinary improvement in clustering performance over choosing a best model. At times, our merging approach requires that components are merged and we introduced methods to do this. We explored two options for determining the number of components to merge over: we can choose the number of components attached to the best model (Case I) or we can choose the model in S with the lowest number of components (Case II).

The approach introduced herein is just the beginning and there are tremendous opportunities to build upon

Table 5: Results for top 5 model selection, merging, and averaging for the simulated data.

	Model Selection			Merging		Average Weight	ARI
	Models	Criteria	G	Case I	Case II		
BIC	1st(EII)	-9753.587	5	no need	$\{1, 2 \cup 3, 4, 5\}$	0.2002342	Best: 0.7135163
	2nd(EVI)	-9763.885	4	no need	no need	0.200023	Avrg. (Case I):
	3rd(EEE)	-9767.062	6	$\{1, 2, 3, 4, 5 \cup 6\}$	$\{1, 2 \cup 4, 3, 5 \cup 6\}$	0.1999579	0.8227513
	4th(EII)	-9768.561	7	$\{1, 2 \cup 3, 4, 5, 6 \cup 7\}$	$\{1, 2 \cup 3 \cup 5, 4, 6 \cup 7\}$	0.1999272	Avrg. (Case II):
	5th(EEI)	-9771.959	7	$\{1, 2 \cup 3, 4, 5, 6 \cup 7\}$	$\{1, 2 \cup 3 \cup 5, 4, 6 \cup 7\}$	0.1998577	0.8777635
ICL	1st(EVI)	-9795.119	4	no need		0.2005361	Best:
	2nd(EEV)	-9812.766	4	no need		0.2001755	0.8760625
	3rd(EEE)	-9822.785	6	$\{1, 2 \cup 4, 3, 5 \cup 6\}$	no need	0.1999713	
	4th(EVI)	-9835.585	5	$\{1, 2, 3, 4 \cup 5\}$		0.1997111	Avrg. (Case I):
	5th(EII)	-9840.764	5	$\{1, 2, 3, 4 \cup 5\}$		0.199606	0.8760625
AIC	1st(EVI)	-9638.797	9		$\{1, 2, 3, 4, 5, 6 \cup 7, 8, 9\}$	0.2000247	Best:
	2nd(EII)	-9639.507	8		no need	0.20001	0.5347199
	3rd(EVI)	-9640.244	8	no need	no need	0.1999947	
	4th(EEI)	-9640.36	8		no need	0.1999923	Avrg. (Case I):
	5th(VEI)	-9641.026	8		no need	0.1999784	0.5931062

Table 6: Top M model selection, merge and average for the simulated data.

	Model Selection			Merge		Average Weight	ARI
	Models	Criteria	Clusters	Case 1	Case 2		
BIC	1st(EII)	-9753.587	5	no need	no need		Best:0.7135163
ICL	1st(EVI)	-9795.119	4	no need	no need		Best:0.8760625
AIC	1st(EVI)	-9638.797	9		$\{1 \cup 9, 2 \cup 3 \cup 5, 4 \cup 8, 6, 7\}$	0.05558314	Best: 0.5347199
	2nd(EII)	-9639.507	8		$\{1, 2 \cup 3 \cup 5, 4 \cup 8, 6, 7\}$	0.05557905	
	3rd(EVI)	-9640.244	8		$\{1, 2 \cup 3 \cup 5, 4 \cup 8, 6, 7\}$	0.0555748	
	4th(EEI)	-9640.36	8		$\{1, 2 \cup 3 \cup 5, 4 \cup 8, 6, 7\}$	0.05557413	
	5th(VEI)	-9641.026	8		$\{1, 2 \cup 3 \cup 5, 4 \cup 8, 6, 7\}$	0.05557029	
	6th(EEE)	-9641.448	6		$\{1, 2 \cup 4, 3, 5, 6\}$	0.05556786	
	7th(EEI)	-9642.014	7		$\{1, 2 \cup 3 \cup 5, 4, 6, 7\}$	0.0555646	
	8th(EVI)	-9642.498	7		$\{1, 2 \cup 3 \cup 5, 4, 6, 7\}$	0.05556181	
	9th(EII)	-9643.443	9	no need	$\{1 \cup 9, 2 \cup 3 \cup 5, 4 \cup 8, 6, 7\}$	0.05555636	Avrg. (Case II): 0.8811323
	10th(EEE)	-9643.724	7		$\{1, 2 \cup 3 \cup 5, 4, 6, 7\}$	0.05555474	
	11th(VEI)	-9644.026	7		$\{1, 2 \cup 3 \cup 5, 4, 6, 7\}$	0.055553	
	12th(EEE)	-9644.5	8		$\{1, 2 \cup 3 \cup 5, 4 \cup 8, 6, 7\}$	0.05555028	
	13th(VII)	-9645.838	9		$\{1 \cup 9, 2 \cup 3 \cup 5, 4 \cup 8, 6, 7\}$	0.05554257	
	14th(EEI)	-9646.4	9		$\{1 \cup 9, 2 \cup 3 \cup 5, 4 \cup 8, 6, 7\}$	0.05553934	
	15th(EEE)	-9646.47	9		$\{1 \cup 9, 2 \cup 3 \cup 5, 4 \cup 8, 6, 7\}$	0.05553893	
	16th(EII)	-9647.279	7		$\{1, 2 \cup 3 \cup 5, 4, 6, 7\}$	0.05553427	
	17th(VVV)	-9648.39	6		$\{1, 2 \cup 4, 3, 5, 6\}$	0.05552788	
	18th(EVI)	-9648.552	5		no need	0.05552695	

it. We illustrated merging for three different criteria (BIC, ICL, and AIC) but the same approach could be used with other criteria (e.g., the LASSO-penalized BIC of Bhattacharya and McNicholas, 2012). We used Gaussian mixture models but the same merging procedures could be applied to non-Gaussian mixtures. We focused on clustering applications, but our merging methods could also be applied for model-based

classification (cf. Dean et al., 2006; McNicholas, 2010). The idea of a top 5 approach, while we have shown it can be effective, is a little *ad hoc* and so is the use of a criteria value difference of 10 in the top M approach. One could, for example populate S with all models with a criterion value within some proportion of the best model. The choice of weight is also flexible. The weight definition used herein achieved improved clustering performance but it needs to be tested extensively to better establish its effectiveness. This, along with everything else mentioned in this paragraph, will form the basis of future work.

Some may criticize this approach because the final result is not an interpretable model in the sense that a Bayesian model averaging might produce an interpretable model. They may go further and argue that the main advantage of mixture model-based clustering is lost when one averages, as we do, placing value on classification accuracy ahead of model interpretability. Of course, as mentioned herein, the idea of reporting clustering results for one model as if the other 89, say, had not been run could also be considered questionable. Whatever about the philosophical debate, if the ultimate goal of clustering is to obtain the best possible estimated group memberships then we maintain that our results show that merging model-based clustering results is worth a further look.

Acknowledgements

This work was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada, the University Research Chair in Computational Statistics at the University of Guelph, and an Early Researcher Award from the Ontario Ministry of Research and Innovation.

References

- Banfield, J. D. and A. E. Raftery (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* 49(3), 803–821.
- Baudry, J.-P., A. E. Raftery, G. Celeux, K. Lo, and R. Gottardo (2008). Combining mixture components for clustering. Research Report RR-6644, INRIA.
- Bhattacharya, S. and P. D. McNicholas (2012). A LASSO-penalized BIC for mixture model selection. arXiv preprint [arXiv:1211.6451](https://arxiv.org/abs/1211.6451).
- Biernacki, C., G. Celeux, and G. Govaert (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(7), 719–725.
- Bozdogan, H. and S. Sclove (1984). Multi-sample cluster analysis using Akaike’s information criterion. *Annals of the Institute of Statistical Mathematics* 36, 163–180.
- Celeux, G. and G. Govaert (1995). Gaussian parsimonious clustering models. *Pattern Recognition* 28(5), 781–793.
- Dasgupta, A. and A. E. Raftery (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association* 93, 294–302.
- Dean, N., T. B. Murphy, and G. Downey (2006). Using unlabelled data to update classification rules with applications in food authenticity studies. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 55(1), 1–14.
- Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B* 39(1), 1–38.

- Forina, M., C. Armanino, M. Castino, and M. Ubigli (1986). Multivariate data analysis as a discriminating method of the origin of wines. *Vitis* 25, 189–201.
- Forina, M. and E. Tiscornia (1982). Pattern recognition methods in the prediction of Italian olive oil origin by their fatty acid content. *Annali di Chimica* 72, 143–155.
- Fraley, C. and A. E. Raftery (1998). How many clusters? Which clustering methods? Answers via model-based cluster analysis. *The Computer Journal* 41(8), 578–588.
- Fraley, C. and A. E. Raftery (2006). MCLUST version 3 for R: Normal mixture modeling and model-based clustering. Technical Report 504, Department of Statistics, University of Washington. Minor revisions January 2007 and November 2007.
- Hennig, C. (2010). Methods for merging Gaussian mixture components. *Advances in Data Analysis and Classification* 4, 3–34.
- Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of Classification* 2, 193–218.
- Kass, R. E. and L. Wasserman (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association* 90(431), 773–795.
- Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā. The Indian Journal of Statistics. Series A* 62(1), 49–66.
- Koehler, A. B. and E. S. Murphree (1988). A comparison of the Akaike and Schwarz criteria for selecting model order. *Journal of the Royal Statistical Society: Series C* 37(2), 187–195.
- Krivitsky, P. N., M. S. Handcock, A. E. Raftery, and P. D. Hoff (2009). Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social Networks* 31(3), 204–213.
- Leroux, B. G. (1992). Mixture models: Theory, geometry and applications. *The Annals of Statistics* 1992, 1350–1360.
- Li, J. (2005). Clustering based on a multilayer mixture model. *Journal of Computational and Graphical Statistics* 14(3), 547–568.
- McLachlan, G. J. and T. Krishnan (2008). *The EM Algorithm and Extensions* (2nd ed.). New York: Wiley.
- McNicholas, P. D. (2010). Model-based classification using latent Gaussian mixture models. *Journal of Statistical Planning and Inference* 140(5), 1175–1181.
- McNicholas, P. D., K. R. Jampani, A. F. McDaid, T. B. Murphy, and L. Banks (2011). *pgmm: Parsimonious Gaussian Mixture Models*. R package version 1.0.
- McNicholas, P. D. and T. B. Murphy (2010). Model-based clustering of microarray expression data via latent Gaussian mixture models. *Bioinformatics* 26(21), 2705–2712.
- Murphy, T. B., N. Dean, and A. E. Raftery (2010). Variable selection and updating in model-based discriminant analysis for high dimensional data with food authenticity applications. *The Annals of Applied Statistics* 4(1), 396–421.
- Qiu, W. and H. Joe (2006). Generation of random clusters with specified degree of separation. *Journal of Classification* 23, 315–334.

- Qiu, W. and H. Joe. (2012). *clusterGeneration: random cluster generation (with specified degree of separation)*. R package version 1.2.9.
- R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66, 846–850.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* 6, 461–464.
- Stanford, D. and A. Raftery (2000). Finding curvilinear features in spatial point patterns: principal curve clustering with noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(6), 601–609.
- Wehrens, R., L. M. Buydens, C. Fraley, and A. E. Raftery (2004). Model-based clustering for image segmentation and large datasets via sampling. *Journal of Classification* 21, 231–253.
- Wolfe, J. H. (1963). Object cluster analysis of social areas. Master’s thesis, University of California, Berkeley.
- Yeung, K. Y., C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics* 17(10), 977–987.